PDBMD2CD: Providing Predicted Protein Circular Dichroism Spectra from Multiple Molecular Dynamics-Generated Protein Structures

Elliot D. Drew and Robert W. Janes*

School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK

*To whom correspondence should be addressed. Tel: +44 207 8828442; Email: r.w.janes@qmul.ac.uk

ABSTRACT

PDBMD2CD is a new webserver capable of predicting circular dichroism (CD) spectra for multiple protein structures derived from molecular dynamics (MD) simulations, enabling predictions from thousands of protein atomic coordinate files (e.g., MD trajectories) and generating spectra for each of these structures provided by the user. Using MD enables exploration of systems that cannot be monitored by direct experimentation. Validation of MD-derived data from these types of trajectories can be difficult via conventional structure-determining techniques such as crystallography or NMR spectroscopy. CD is an experimental technique that can provide protein structure information from such conditions. The website utilises a much faster (minimum ~1000x) and more accurate approach for calculating CD spectra than its predecessor, PDB2CD (1). As well as improving on the speed and accuracy of current methods, new analysis tools are provided to cluster predictions or compare them against experimental CD spectra. By identifying a subset of the closest predicted CD spectra derived from PDBMD2CD to an experimental spectrum, the associated cluster of structures could be representative of those found under the conditions in which the MD studies were undertaken, thereby offering an analytical insight into the results. PDBMD2CD is freely available at: https://pdbmd2cd.cryst.bbk.ac.uk.

INTRODUCTION

Circular Dichroism (CD) spectroscopy is a widely used technique to explore and examine aspects of protein structure in solution. CD data can be employed to determine the secondary structure content of a protein and hence, whether a protein is correctly folded or not. Notably, it is also a dynamic technique able to monitor any resultant structural changes that arise when ambient solution conditions are altered, for example by temperature or pH. Additionally, identifying changes in structure following the binding of ligands, cofactors, or other proteins is also possible. So the technique offers a highly

versatile approach to monitoring the dynamics of structural changes of proteins in solution when it is not possible via other techniques such as crystallography, or even solution nuclear magnetic resonance (NMR) studies, to make such comparisons.

Molecular Dynamics (MD) simulation studies, specifically those on proteins, can be used to examine the structure and dynamics of complex macromolecular systems as they evolve over time. As the name suggests, these are *in silico* simulations capable of offering atomistic and dynamic insight on systems, even where there are no experimental techniques available to obtain such data directly. In these studies, it is important to validate the MD results whenever possible, by examining the extent of the match between data generated computationally and any comparable data obtained from a supporting experimental technique. Data that match well lends weight, and hence validity, to other properties obtained from an MD study where those specific results cannot be supported by direct experimental means. The use of MD for studying interactions involving proteins is growing year on year, so it is all the more important to have an experimental technique which is versatile enough to provide the necessary supporting data to match with the MD results. CD spectroscopy is readily available as such a technique.

Previously we developed a webserver, PDB2CD (1), dedicated to producing predictions of circular dichroism spectra from protein atomic coordinates. That package has fulfilled an important role in providing users with meaningful predicted CD spectra of proteins where obtaining an experimental CD spectrum was not possible, often enabling comparison of this against an actual spectrum from a related protein (2-6).

However, PDB2CD was not envisioned to cater for multiple input files due to the inherent "rate determining step", that of performing the structure comparisons between the query structure and those in the reference set of proteins used, the SP175 (7) "gold standard" set available in the Protein Circular Dichroism Data Bank (PCDDB) (8). This meant that to generate predicted CD spectra using PDB2CD, structures derived in an NMR ensemble had to be input individually. For structures generated from MD studies, only the average, or most representative structure from the trajectories would be entered (5,6)). To provide a site with additional analysis tools capable of dealing with generating predicted CD spectra from multiple protein structure input files in an efficient and accurate way, the PDBMD2CD webserver has been created.

THE PDBMD2CD SERVER

The PDBMD2CD webserver is a user-friendly resource designed with the aims of making it available to a wider user community. Specifically, the focus has been to facilitate the input of multiple protein structure entries from which predicted CD spectra are generated for each one. Our aims were to optimise on speed of delivering the results whilst retaining, and improving, the accuracy levels of CD spectra prediction. Enabling multiple structure input means that all selected structures sampled from MD trajectories as a representative set can be used to generate information over the entire time-course of the study, rather than using only one "representative" structure (5,6).

An overview of the methodology behind the PDBMD2CD package is presented here, with more details in the Supplementary Data. PDBMD2CD employs a combination of two different approaches: The first

uses basis spectra that represent seven secondary structural types derived from a least squares regression of the CD signal (Figure S1 and Table S2) from structures in our 83 protein reference set (a set derived from the high quality SMP180 (9) which has a broader range of proteins incorporating both soluble and membrane proteins, and shown in Table S1). The second creates a basis set of spectra from structures in the reference set with the closest secondary structure content to the query protein, estimating each basis spectra's weighting in the final prediction through a multivariate optimisation of the contributing spectra's summed secondary structure content. The two predicted spectra from these approaches are then averaged to give the final prediction, resulting in better accuracy than either method alone.

Testing of PDBMD2CD was performed using leave-one-out (LOO) cross-validation on the training set using the Root Mean Square Deviation (RMSD) between predicted and experimental spectra as a performance metric. In addition, predictions were made on a separate test set (8 structures derived from the PCDDB) was used (Tables 1 and S3). Summarised data comparisons are presented for PDB2CD (1), the DichroCalc webserver (10) and SESCA (for the test set only), (11), a downloadable, command-line python package. Results from PDBMD2CD (Table 1) were in good agreement overall with the experimental spectra and showed significant improvements over existing *ab initio* and empirical methods, including PDB2CD. Figure 1 shows comparisons of the RMSD differences between the experimental and predicted CD spectra for PDBMD2CD, PDB2CD and DichroCalc, ordered from the lowest to highest RMSD values both for the LOO cross-validation results and for the test set of proteins. Examples of predicted CD spectra for the test set structures are presented in Figure 2. Two of the better-predicted spectra are shown in Panels A and B, whilst two of the poorer predictions are shown in Panels C and D.

INPUT AND OUTPUT

Input

PDBMD2CD takes as input Protein Data Bank (PDB) structure files, either in PDB or mmCIF format. Single or multiple structure files may be uploaded. To minimise wait time caused by upload speeds, large numbers of files can be uploaded as a single compressed (zip, bzip or tar.gz archive) file. Alternatively, it is possible to use a PDB code, or multiple codes separated by commas, placed into the box provided, so the structure files corresponding to those codes are fetched from the RCSB PDB (12) servers. Although the main focus has been on files from MD simulations, multiple files from other sources can also be used. For example, these could be from an NMR ensemble of structures where it would be of interest to see the range of conformations obtained, particularly where flexibility in the structure might lead to differing clusters of structures; CD could be used as to arbitrate between these possible outcomes. Another illustration of potential usage is in homology studies, taking the structures of a range of related proteins from the PDB and then comparing their predicted spectra to that of an experimental one to ascertain which of the homologous structures most resembles that of the experimental protein.

Output

PDBMD2CD creates three output pages which may be accessed through a tabbed menu. These are the **Results**, which is the default page, **Clustering**, and **Compare to Experiment** pages. Each page has specific types of analyses for the multiple predicted spectra. An example of how these pages appear, together with the home webpage, is given in Figure 3 and is discussed in the "Case Study" section.

Results tab

On this page are all the generated CD spectra together with the mean of these spectra in one interactive graphic. To the right of that is an interactive graphics pane showing the structure most representative of the set of structures entered. Below this is a secondary structure plot of alpha helix versus beta sheet content derived as derived by the "Dictionary of Secondary Structure of Proteins" (DSSP) definition (13), showing in blue the data for the reference set of structures used and in orange the positions of each of the structures uploaded for generating their CD spectra. To the left of this are data associated with the generated CD spectra: the names of the structures used in the predictions, the date and time of the run, the number of structures uploaded, a report on the number of any failures within this set (with a pull down box below this value indicating which these are if there are any), the average RMSD of the generated spectra from the mean of the set, and the name of the most representative structure file (the structure shown in the upper right pane) in PDB format. There are links available to this file in this table and also below the interactive structure pane. The spectra are downloadable both as .zip and .csv files. Below this are the mean values for the secondary structure content present in the structures, with an associated standard deviation. These secondary structure features are: Helix 1, Helix 2, Antiparallel Sheet 1, Antiparallel Sheet 2, Parallel Sheet, Turn, and Other (defined as the content of what is left that is not defined by those terms above it) such that the total adds to 100% (defined in the Supplementary Data).

Clustering tab

On this page it is possible to group the predicted spectra into clusters using the k-means clustering (14) method (see Supplementary Data for the full explanation of this approach). Once again, the overall set of predicted spectra are presented on the top left of the page. On the right of this is an "elbow" plot, which allows visual estimation of the optimal number of clusters within this set of spectra by identifying the smallest number of clusters which account for the largest variation in the data. The elbow plot shows the distortion, calculated as the mean Euclidean distance between the cluster members and the centroids generated for values of k from 1 through to 6 inclusive. The largest deviation from linearity between three contiguous k-mean cluster values will produce a bend forming the "elbow" and the middle value of the three represents the most likely number of clusters present in the data set (see also the Supplementary Data). The plot gives a guide to the number of clusters present in the data, but the user needs to make the ultimate judgement on what might be considered the best number of clusters.

As the k value is modified by the user (from a pull down box at the top of the Clustering page), so the spectra within the main representation are coloured according to the numbers of clusters identified, and below this main plot are the individual clusters similarly coloured. Each cluster has separate data

associated with them which includes to the right of the plot the average RMSD of the clustered spectra from its mean, the identity (and a link to it) of the representative structure of that cluster, the number of structures within that cluster, the average secondary structure content, and a download link to the spectra associated within the cluster, in .csv format. To the right of this is an interactive pane displaying the most representative structure of that cluster.

Compare to Experiment tab

This page enables comparisons to be made between the predicted spectra and a user-provided experimentally-derived spectrum. The user file can be uploaded in many standard text-based output formats (produced by different CD instruments), as an output file from the Synchrotron Radiation CD (SRCD) beamlines, and as a generic two-column format (wavelength versus CD data). Uploaded spectra can be in either Delta Epsilon units (as default) or in Mean Residue Ellipticity (chosen by selecting a radio button), which in this case the package converts into Delta Epsilon.

The RMSDs between the experimental and each of the predicted CD spectra are used as the measure for comparison. The user may choose between comparing using an RMSD threshold or by specifying the number of predicted spectra to show that are closest to the experimental spectrum. By default, this RMSD value (which may be modified by the user), is initially set at 0.5, or to half the maximum RMSD value if that value is smaller than 0.5. As the user choses a maximum RMSD value so the value showing of the number of closest predicted spectra is updated to that which equates to the chosen RMSD value, and the reciprocal of this happens if a number of closest spectra is chosen instead. Each comparison will generate a subset of the predicted spectra as a result. To the left below this section is a plot displaying the currently chosen subset of spectra together with the experimental spectrum in red, the mean prediction of the subset in blue, and the closest predicted spectrum to the experimental spectrum in green. To the right of this is a pane showing a histogram plot of the distribution of predicted spectra RMSD values from the experimental. This can be displayed in either of two forms: the counts of spectra as binned blocks of the RMSD of the predicted spectra from the experimental spectrum, or the cumulative sum of these counts. A solid red line indicates the position of the maximum RMSD value in the currently selected subset. Moving the cursor over the plot generates a dashed red line which moves with it displaying as it does so, the values of RMSD, count and cumulative count, allowing a user to select a new threshold RMSD value for subset selection. Clicking the left mouse button generates the new comparison position and the solid red line moves to that new position. Below this pane are detailed the experimental file name, the number of members in the current displayed subset of spectra, the matching maximum RMSD threshold for that subset, the mean RMSD of the subset, the name of the structure file within the subset from which the closest predicted spectrum to the experimental is derived, the closest RMSD of this spectrum, and the structure furthest away in this subset, and its related RMSD value. To the left of this data, and below the plot, is the mean 7-state secondary structure information for the current subset as a percentage, with its associated standard deviation. The names and predicted spectra of the structures in the subset can be downloaded as a .csv file.

CASE STUDY – UNFOLDING SIMULATIONS OF HEN EGG-WHITE LYSOZYME

To illustrate and highlight the way in which this webserver might be used to provide valuable information and analyses of data, an MD simulation experiment was undertaken studying the thermal unfolding and refolding properties of Hen egg-white lysozyme (HEWL). To generate the spectra a zipped file containing the MD structures was uploaded (Figure 3A). The predicted spectra were displayed on the Results page together with the most representative structure, together with data pertaining to the input files and secondary structure information on these structures (Figure 3B). Clustering of these predicted spectra, together with associated data on these clusters, were produced (Figure 3C) and comparisons were obtained to a series of individual experimental SRCD spectra (Figure 3D).

A similar MD study by Meersman et al, (15) used pooled experimental data from SRCD spectroscopy, Fourier Transform Infra-red (FTIR) spectroscopy, NMR, small-angle X-ray scattering (SAXS) studies and MD simulations to examine the effects of temperature on HEWL. In this earlier study all these respective experimental techniques obtained results over the temperature range of 20 ^{O}C to 77 ^{O}C, and to match this the maximum time-course for the MD runs was set at 10 nanoseconds (ns), and repeated nine times, at a temperature of 500 K. Here a different strategy was adopted; conducting MD runs as two repeats of 500 ns at temperatures of 270, 300, 350 and 450 K (full details of the MD simulation protocol are given in the Supplementary Data). Reference set structures were recorded for each trajectory; the first structure at t=0, and at every 1.5 ns thereafter, such that a total of 2672 structures were produced. All these structures were pooled into one data set. The reasoning for this strategy was that as the range of temperatures of the MD runs was broad the structures produced would be most reflective of the folded state in the lowest temperature studies, would match the partially folded states in the mid temperature runs, and would match the unfolded state in the higher temperature runs.

To ascertain the quality and relevance of the MD structures produced from our study, parameters such as secondary structure content, and radius of gyration, obtained experimentally from the other techniques reported in the original paper (15) were also generated from the structures in our MD simulations. Our criterion for MD structures to be selected as representative of the temperatures over the experimental range was solely based on the degree of matching of their predicted CD spectra to those experimentally spectra reported in the SRCD data. These spectra were obtained from the Protein Circular Dichroism Data Bank (PCDDB) (8) resource under entry codes CD0003675000.gen to CD0003675013.gen for the unfolding spectra, and code CD0003690000.gen for the refolded from 72 ^{O}C and that of the refolded 77 ^{O}C data (A.J. Miles, personal communication). Each experimental SRCD spectrum from the study was compared in turn to the 2672 predicted CD spectra generated by PDBMD2CD. For each the ensemble group size of closest predicted spectra (and hence of the associated MD structures) was taken to be 60, (just over 2% of the total predicted spectra) thereby keeping the data close and relevant to the results.

Table 2 shows the comparisons between the experimental data obtained from the melting studies of lysozyme and the same parameters generated from the MD structures. The experimental CD data show lysozyme to be a very resilient structure to thermal unfolding. Little change in helix content (4%) is seen between 20 and 64 ^{O}C (helix % (CD) in Table 2). Above this temperature the helix loss is more pronounced, where at 77 ^{O}C the calculated content is around 19%. It is possible other forms of “structure” are present at these higher temperatures which cannot be determined from the CD data.

This is because there is an increase in the Normalised RMSD (NRMSD) value compared to those of the lower temperature values. This term is a "goodness-of-fit" between the experimental spectrum and the spectrum back-calculated using the calculated secondary structure content, which should be as close to zero as possible. The term comes from the CONTINLL (16) method that was used in DichroWeb (17) to obtain the helix secondary structure content of the experimental CD spectra. The helix content calculated from the 60 closest MD structures (Helix % (MD) in Table 2) for each of the temperatures show an excellent agreement with the CD values over the entire temperature range to 72 ^{O}C. Only the 77 ^{O}C values differ substantially where, as stated, the analysis of the CD data may have some issues as shown by a poor NRMSD value.

Figure 4 shows representative MD simulation structures for selected temperatures from over the range of the study (the full set of structures with further associated information is given in the Supplementary Data (Figure S2)). The radii of gyration values obtained from the SAX studies (Rg (SAXS) in Table 2) also show excellent agreement with those produced from each of these representative groups of MD structures. These indicate, again, that up to 64 to 68 ^{O}C there is little structural change in the volume of lysozyme; i.e. the tertiary structure remains intact whilst only the ends of some helices start to unfold (as shown in Figure 4). At the highest temperature only the core of one helix and some beta sheet structure remain. Transient helix propensities are retained sufficiently at 72 ^{O}C to provide a nucleus to enable recovery of the tertiary structure on cooling to quite a reasonable extent, as indicated by the data associated with the cooling back to 20 ^{O}C from 72 ^{O}C (R72 in Table 2). At 77 ^{O}C these core helices are now lost to the point that their propensities are so low that it is not possible to recover the tertiary structure when the protein is cooled back to 20 ^{O}C as indicated by the R77 ^{O}C data (Table 2).

SUMMARY

PDBMD2CD provides the user with a means of obtaining predicted CD spectra from multiple input coordinate files from a variety of sources; MD simulation structures, NMR ensemble structures, and multiple homologous proteins, for example. The webserver provides a ready means of interrogating predicted spectra in terms of possible clustering, offering insight into potentially different populations of structures present in the original input, or through a direct comparison to an experimental spectrum, perhaps indicating that a group of the input structures have characteristics comparable to those of the protein that produced that experimental data. As illustrated in the lysozyme case study, PDBMD2CD offers a ready means of analysing the structures from MD simulations grouping these purely by matching the degree of similarity between their predicted CD spectra to those of an experimentally-derived set of spectra from thermal unfolding and refolding studies. The helical content, and radii of gyration generated from these groups showed excellent agreement to the comparable experimentally determined values across a wide range of temperatures. The information and analysis potential demonstrated by the PDBMD2CD webserver shows that this fast and user-friendly site can provide a novel way to interrogate MD structural data.

SUPPLEMENTARY DATA

Supplementary data are available through NAR online.

ACKNOWLEDGEMENTS

We thank Prof. B.A. Wallace (Birkbeck College, University of London) for invaluable discussions regarding the manuscript and her and her group for providing useful feedback in the testing of the package during development. We also particularly thank Dr. Andrew J. Miles for providing one of the spectra not in the PCDDB, namely the R77 refold data not available in the PCDDB.

FUNDING

Biotechnology and Biological Sciences Research Council (BBSRC), U.K, [BB/J019194, BB/P024106 to R.W.J.]. Funding for open access charge is from UKRI through the BBSRC [BB/J019194, BB/P024106].

Conflict of interest statement. None declared.

REFERENCES

1. Mavridis,L. and Janes,R.W. (2017) PDB2CD: a web-based application for the generation of circular dichroism spectra from protein atomic coordinates. *Bioinformatics*, 33, 56-63. https://doi.org/10.1093/bioinformatics/btw554

2. Mendes,L.F.S., Fontana,N.A., Oliveira,C.G., Freire,M.C.L.C., Lopes,J.L.S., Melo,F.A. and Costa-Filho A.J. (2019) The GRASP domain in golgi reassembly and stacking proteins: differences and similarities between lower and higher Eukaryotes. *FEBS J.*, 286, 3340-3358. https://doi.org/10.1111/febs.14869

3. Caveney,N.A., Pavlin,A., Caballero,G., Bahun,M., Hodnik,V., de Castro,L., Fornelos,N., Butala,M. and Strynadka,N.C.J. (2019) Structural Insights into Bacteriophage GIL01 gp7 Inhibition of Host LexA Repressor. *Structure,* 27,1094-1102. https://doi.org/10.1016/j.str.2019.03.019

4. Osterlund,N., Kulkarni,Y.S., Misiaszek,A.D., Wallin,C., Kruger,D.M., Liao,Q.H., Rad,F.M., Jarvet,J., Strodel,B. and Warmlander,S.K.T.S. (2018) Amyloid-β Peptide Interactions with Amphiphilic Surfactants: Electrostatic and Hydrophobic Effects. *ACS Chem. Neurosci*., 9, 1680-1692. https://doi.org/10.1021/acschemneuro.8b00065

5. Wang,F., Orioli,S., Ianeselli,A., Spagnolli,G., Beccara,S.A., Gershenson,A., Faccioli,P. and Wintrode,P.L. (2018) All-Atom Simulations Reveal How Single-Point Mutations Promote Serpin Misfolding. *Biophysical J.,* 114, 2083-2094. https://doi.org/10.1016/j.bpj.2018.03.027

6. Zheng,X., Mueller,G.A., Kim,K., Perera,P., Eugene,F., DeRose,E.F. and London,R.E. (2017) Identification of drivers for the metamorphic transition of HIV-1 reverse transcriptase. *Biochem J.*, 474, 3321-3338. https://doi.org/10.1042/BCJ20170480

7. Lees,J.G., Miles,A.J., Wien,F. and Wallace,B.A. (2006) A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics,* 22, 1955-1962. http://www.ncbi.nlm.nih.gov/pubmed/16787970

8. Whitmore,L., Miles,A.J., Mavridis,L., Janes,R.W. and Wallace,B.A. (2017) PCDDB: new developments at the Protein Circular Dichroism Data Bank. *Nucleic Acids Res*., 45(D1), D303-D307. http://nar.oxfordjournals.org/content/45/D1/D303

9. Abdul-Gader,A., Miles,A.J. and Wallace,B.A. (2011) A reference dataset for the analyses of membrane protein secondary structures and transmembrane residues using circular dichroism spectroscopy. *Bioinformatics,* 27, 1630-1636. https://doi.org/10.1093/bioinformatics/btr234

10. Jasim,S.B., Li,Z., Guest,E.E. and Hirst,J.D. (2018) DichroCalc: Improvements in Computing Protein Circular Dichroism Spectroscopy in the Near-Ultraviolet. *J. Mol. Biol*., 430, 2196-2202. https://doi.org/10.1016/j.jmb.2017.12.009

11. Nagy,G., Igaev,M., Jones,N.C., Hoffmann,S.V. and Grubmüller.H. (2019) SESCA: Predicting Circular Dichroism Spectra from Protein Molecular Structures. *J. Chem. Theory Comput*., 15, 5087-5102. http://dx.doi.org/10.1021/acs.jctc.9b00203

12. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., 28, 235-242. https://dx.doi.org/10.1093/nar/28.1.235

13. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*., 22, 2577-637. doi:10.1002/bip.360221211. PMID 6667333.

14. MacQueen, J., 1967, June. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

15. Meersman,F., Atilgan,C., Miles,A.J., Bader,R., Shang,W.F., Matagne,A., Wallace,B.A. and Koch,M.H.J. (2010) Consistent Picture of the Reversible Thermal Unfolding of Hen Egg-White Lysozyme from Experiment and Molecular Dynamics. *Biophysical J.,* 99, 2255-2263. https://doi.org/10.1016/j.bpj.2010.07.060

16. Van Stokkum,I.H.M., Spoelder,H.J.W., Bloemendal,M., Van Grondelle,R. and Groen,F.C.A. (1990) Estimation of protein secondary structure and error analysis from CD spectra. *Anal. Biochem*. 191, 110-118. http://doi.org/10.1016/0003-2697(90)90396-q

17. Whitmore,L. and Wallace,B.A. (2008) Protein Secondary Structure Analyses from Circular Dichroism Spectroscopy: Methods and Reference Databases. *Biopolymers,* 89, 392-400. (PDF)

18. Micsonai,A., Wien,F., Bulyáki,E., Kun,J., Moussong,E., Lee,Y.H., Goto,Y., Réfrégiers,M. and Kardos,J. (2015) Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy, *Proc. Nat. Acad. Sci.*, 112, E3095-E3103. doi/10.1073/pnas.1500851112

Table 1: Left panel, cross-validation leave-one-out (LOO) data for the 83 proteins in the reference set and comparison of predicted and right panel, experimental CD spectra of 8 test proteins. The average RMSD values for PDBMD2CD, PDB2CD and DichroCalc for the LOO data, and included is SESCA for the test set of proteins (lowest values are the best and highlighted in **bold**).

LOO Reference set	**RMSD**
PDBMD2CD	**0.940 (0.394)**
PDB2CD (SMP180)	1.026 (0.468)
DICHROCALC	2.062 (1.111)

Test Set (8 Spectra)	**RMSD**
PDBMD2CD	**0.962 (0.496)**
PDB2CD (SMP180)	1.347 (0.443)
DICHROCALC	2.111 (1.596)
SESCA – (DSSP-1 basis set)	1.055 (0.442)

Table 2. Data associated with the Case Study thermal melting of Lysozyme by MD simulations in this paper and experimental data from SRCD and SAXS, obtained from ref 15. The columns are: the temperatures of the SRCD runs (^{O}C), the NRMSD values giving the match between the experimental CD spectrum and the back-calculated spectrum derived from the calculated secondary structure content for each temperature, the RMSD difference between the experimental CD spectra and the closest group of 60 predicted spectra (from the MD structures) at each temperature, the helix content determined from each MD group, the helix content determined from the SRCD spectra by CONTINLL method in the DichroWeb server, the radii of gyration of each group of structures, and the radii of gyration determined experimentally from low-angle X-ray scattering (SAX) studies.

Temperature ^{O}C	**NRMSD (CD)**	**RMSD (CD-MD)**	**Helix % (MD)**	**Helix % (CD)**	**Rg (Å) (MD)**	**Rg (Å) (SAX)**
20	0.021	0.248	40.6 (1.3)	39 (1)	14.2 (0.1)	14.5
24	0.023	0.257	40.5 (1.2)	39 (1)	14.2 (0.1)	-
28	0.024	0.262	40.5 (1.2)	39 (2)	14.2 (0.1)	-
33	0.02	0.257	40.4 (1.2)	39 (2)	14.2 (0.1)	-
37	0.035	0.25	40.4 (1.2)	38 (1)	14.2 (0.1)	-
42	0.028	0.214	39.5 (1.4)	38 (2)	14.2 (0.1)	14.3
47	0.031	0.24	39.7 (1.4)	37 (2)	14.2 (0.1)	-
51	0.035	0.203	38.6 (1.2)	37 (2)	14.2 (0.1)	-
55	0.043	0.2	38.0 (1.3)	37 (1)	14.2 (0.1)	-
60	0.055	0.188	37.2 (1.2)	35 (1)	14.2 (0.1)	14.2
64	0.065	0.189	35.8 (1.5)	35 (0)	14.2 (0.1)	14.3
68	0.075	0.175	33.9 (1.7)	31 (1)	14.1 (0.1)	14.3
72	0.142	0.342	21.1 (7.8)	25 (0)	15.8 (1.8)	14.9
77	0.096	0.513	7.7 (2.9)	19 (1)	16.6 (1.1)	16.6
R72*	0.037	0.215	38.1 (1.2)	36 (1)	14.2 (0.1)	-
R77*	0.157	0.38	16.1 (5.7)	28 (1)	16.7 (2.1)	-

* R72 and R77 refer to temperatures at 20 ^{O}C having returned from 72 and 77 ^{O}C

Figure 1. Plots of the RMSD values between calculated spectra for DichroCalc (dotted line), PDB2CD (thick dashed line), SESCA (right hand panel thin dashed line) and PDBMD2CD (solid line) for the Reference proteins in the leave-one-out cross-validation, (left hand panel) and for the test proteins (right hand panel).

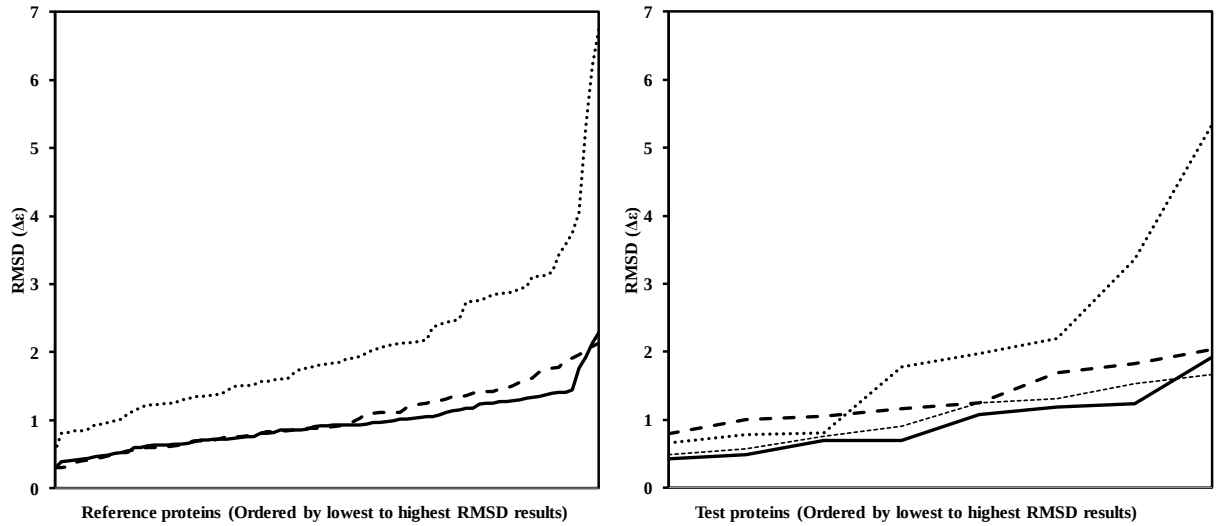


Figure 2. Examples of predicted spectra from DichroCalc (dotted line), PDB2CD (dot-dash line) and PDBMD2CD (dashed line) plotted against the experimentally-determined spectrum (18) (solid line) obtained from the PCDDB (8). Two good PDBMD2CD examples are: A: human dUTPase (PDB Code: 1Q5U) and B: 3-isopropylmalate dehydrogenase (PDB Code: 2Y3Z). Two poorer examples are: C: Ecotin (PDB Code: 1ECZ) and D: Beta-2-microglobulin (PDB Code: 2YXF).

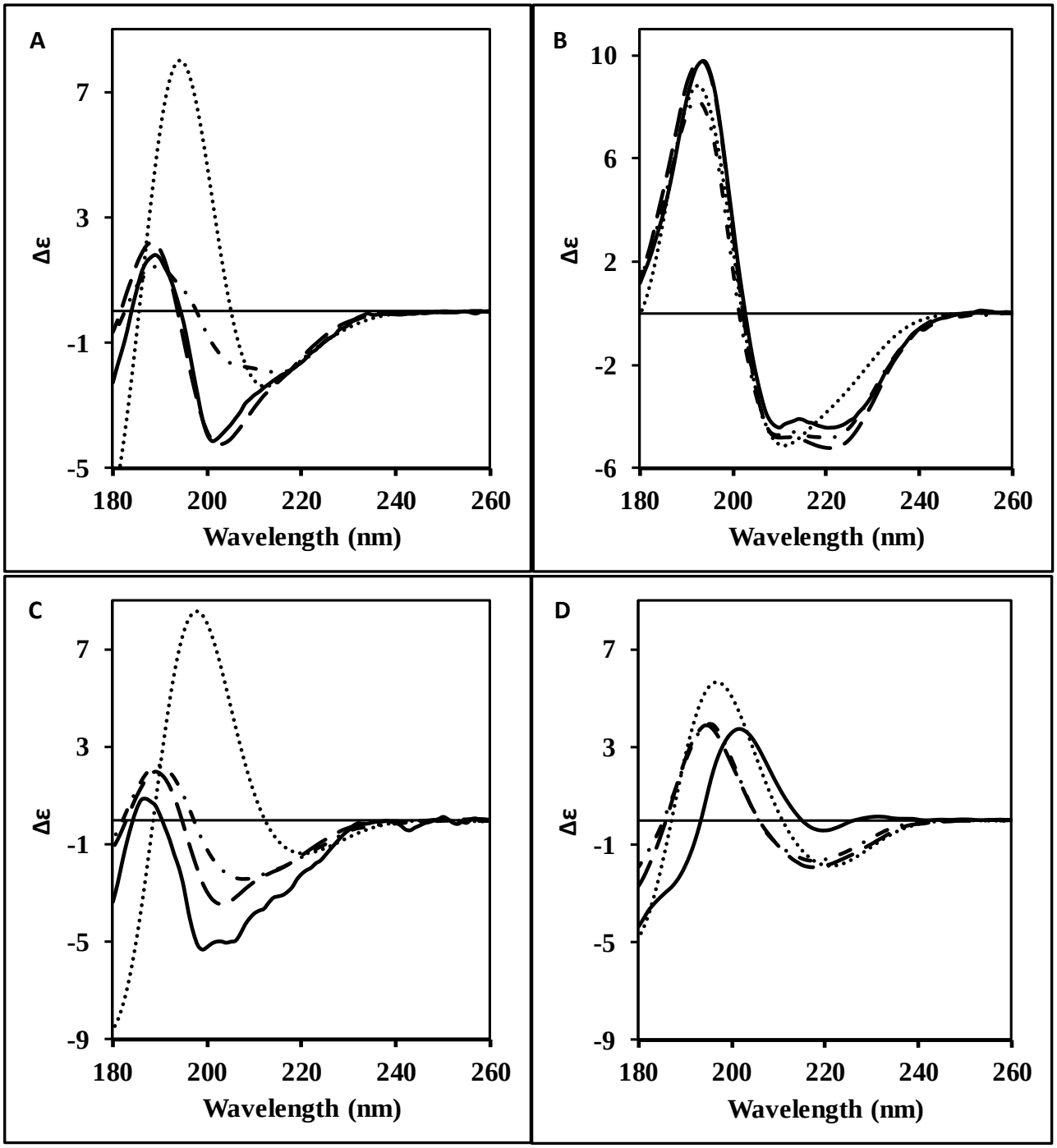


Figure 3.

Example pages from the PDBMD2CD webserver. Panel A is the landing input page; panel B the Results page showing all the predicted spectra (in light blue in the left pane together with the most representative spectrum of the set in dark blue), the most representative structure of the set (in the right upper pane), the secondary structure alpha helix versus beta sheet content in the right hand lower pane, the blue being for the reference set, the orange for the input structures (2672 here), and associated secondary structure information (in the lower left pane); panel C is the Clustering page illustrating the results from a k-mean=2 clustering (shown in the right hand pane), where, below this, the spectra have been clustered into the purple first cluster set, and the dark orange second cluster set, and there are associated secondary structure information and the most representative structures associated with each cluster also shown in the panes to the right of the plots; Panel D is the Compare to Experiment page and shows the information obtainable from comparison to an input experimental spectrum. Here a subset of spectra is being shown with an RMSD smaller or equal to 0.5 as maximum away from the experimental spectrum. The spectra are shown in the left pane in light blue, with the experimental spectrum in red, the subset mean in dark blue, and the closest spectrum to the experimental in green. A histogram plot showing the RMSD distribution of predicted spectra distanced from the experimental spectrum is shown in the right hand pane with a red line indicating the RMSD/number of chosen spectra in the current subset. Below this are secondary structure and associated data pertinent to the current chosen subset.

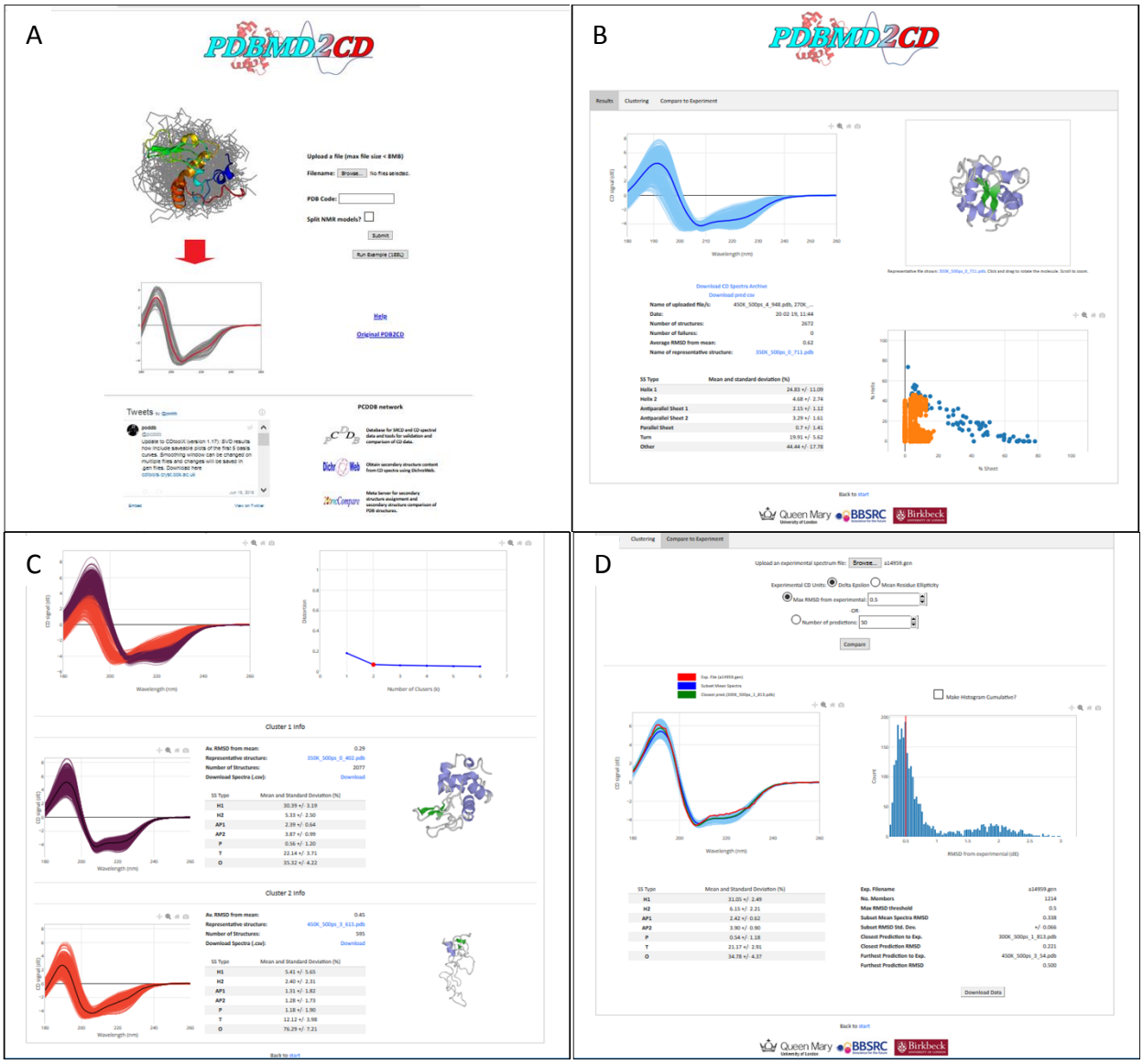
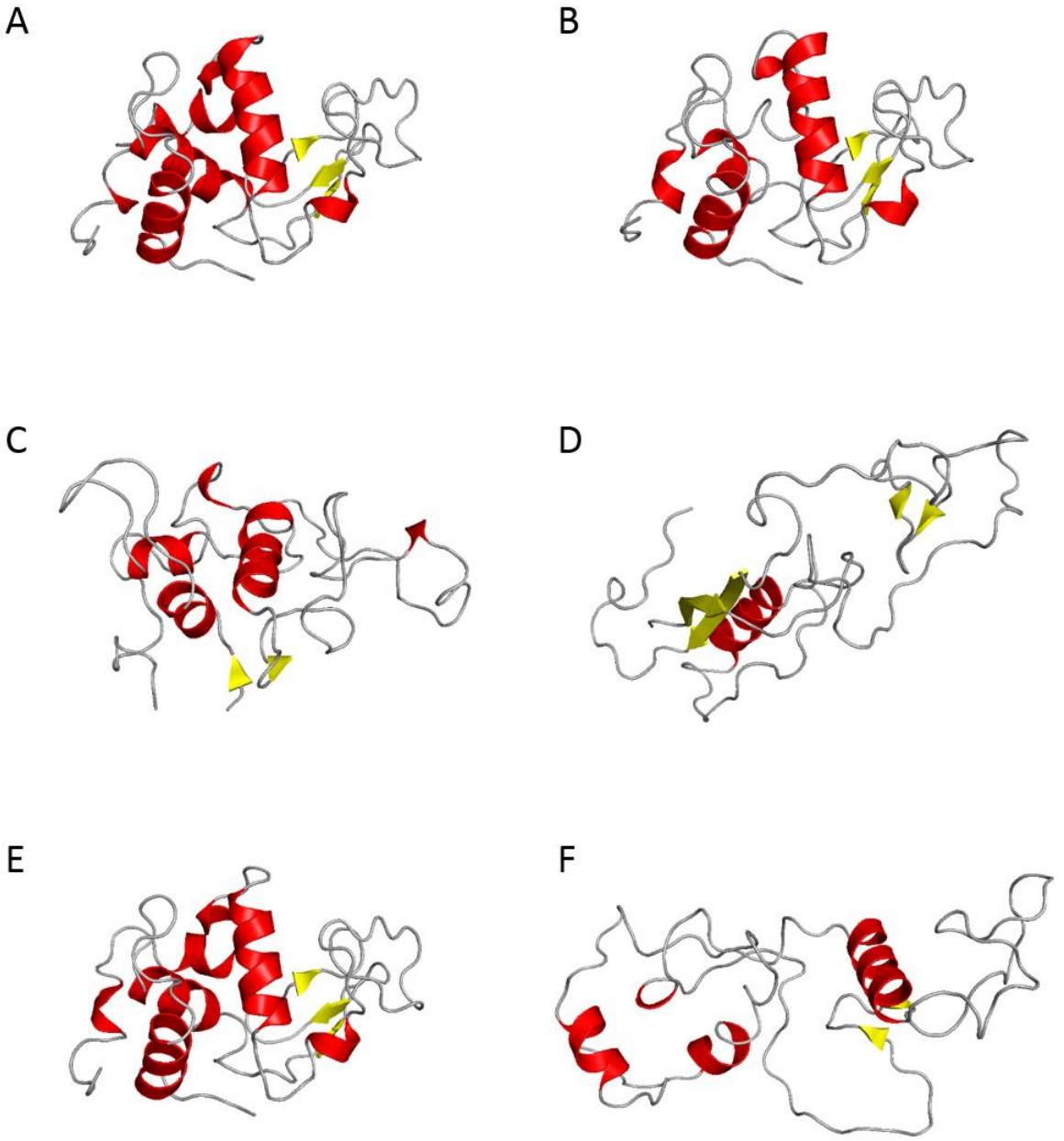

A
PDBMD2CD
Upload a file (max file size = 8MB)
Filename:
Browse...
No files selected.
PDB Code:
Split NMR models?
Submit
Run Example (1BBL)
Help
Original PDB2CD
Tweets
PCDDB network
B
PDBMD2CD
Results
Clustering
Compare to Experiment
Download CD Spectra Archive
Download pred.csv
Name of uploaded file/s:
Date:
Number of structures:
Number of failures:
Average RMSD from mean:
Name of representative structure:
SS Type
Mean and standard deviation (%)
Helix 1
Helix 2
Antiparallel Sheet 1
Antiparallel Sheet 2
Parallel Sheet
Turn
Other
Back to start
Queen Mary
BBSRC
Birkbeck
C
Cluster 1 info
Cluster 2 info
Av. RMSD from mean:
Representative structure:
Number of Structures:
Download Spectra (.csv):
SS Type
Mean and Standard Deviation (%)
H1
H2
AP1
AP2
P
T
O
Back to start
D
Clustering
Compare to Experiment
Upload an experimental spectrum file:
Browse...
Experimental CD Units:
Delta Epsilon
Mean Residue Ellipticity
Max RMSD from experimental:
OR
Number of predictions:
Compare
Make Histogram Cumulative?
SS Type
Mean and Standard Deviation (%)
Exp. Filename
No. Members
Max RMSD threshold
Subset Mean Spectra RMSD
Subset RMSD Std. Dev.
Closest Prediction to Exp.
Closest Prediction RMSD
Furthest Prediction to Exp.
Furthest Prediction RMSD
Download Data
Back to start
Queen Mary
BBSRC
Birkbeck

Figure 4.

Representative structures from the MD simulation studies for the folded/unfolded states at 20, 64, 72, 77 ^{O}C, and recovery to 20 from 72 ^{O}C (R72) and to 20 from 77 ^{O}C (R77) in panels A to F respectively. Where feasible, the orientations of the structures are maintained in each panel. The vertical helix in panel A is that which is retained the most in each of the subsequent panels.



PDBMD2CD: Providing Predicted Protein Circular Dichroism Spectra from Multiple Molecular Dynamics-Generated Protein Structures

Elliot D. Drew and Robert W. Janes*

School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK

*To whom correspondence should be addressed. Tel: +44 207 8828442; Email: r.w.janes@qmul.ac.uk

Supplementary Data

METHODS AND MATERIALS

Reference Dataset

The reference data set consists of 83 proteins (Table S1) selected from the SMP180 “golden standard” SRCD spectral data set (1), used in the DichroWeb deconvolution server (2) and available at the PCDDB (3). All proteins in SMP180 have well-determined SRCD spectra as reported by ValiDichro (4) and have associated PDB structures, and includes the SP175 (5) data set. Proteins with spectral characteristics arising from non-secondary structural phenomena, (for example ligands, or possible exciton-coupling of aromatic side chains), were removed from the data set as these features led to inaccuracy in the predictions produced. In addition, proteins whose removal from the reference set resulted in an increased accuracy during leave-one-out cross validation were ultimately excluded from the 83 proteins of the final reference set. The reasoning for this was that their inclusion in the training set would lead to a decrease in prediction performance for the majority of proteins.

Test data

Eight proteins were identified in the PCDDB which satisfied the criteria of having well-determined CD spectra, as shown by their ValiDichro reports, and having associated PDB structures (Table S3). These were used as a wholly independent further test set for the PDBMD2CD method.

Input to the server

The server accepts both PDB/mmCIF structure files and PDB codes. For analysis of PDB structure files, single files or multiple files can be uploaded. Multiple files can also be uploaded as a zip, bzip or tar.gz archive file. For PDB codes, multiple files can be submitted by separating codes with a comma.

The server has been tested with >1000 structures in a single job, which yielded a result (not including time to upload) in under 5 minutes, making it suitable for MD trajectory analysis.

Secondary structure assignment

The 8 state secondary structure assignments of input structures are calculated by DSSP (6). In Table S2 the mapping between our classification and DSSP classes is shown. Additionally, information about missing residues in the structure (those present in the sample but whose positions could not be assigned often due to intrinsic disorder or high flexibility), is obtained, if present from the header of the structure file. The number of missing residues found is added to the C (considered as "O" here for "Other") DSSP class. The assignments for Beta strand ("E") are further processed to produce the final assignments used in the method.

The CD signals produced by beta-sheet structures are diverse due to the variety of topological arrangements possible in this class. Strands can be arranged in antiparallel or parallel sheets and these can display varying degrees of distortion or twist. Manavalan and Johnson (7) observed that beta-rich proteins fall into two classes with respect to their CD spectra: one class has a "classical" beta-sheet spectrum (negative band ~218 nm, positive band ~195 nm); the other class has a CD spectrum more like that of unordered proteins (negative band near 200 nm). Wu et al. (8) designated these two classes as beta-I and beta-II, respectively. More recently, a treatment of protein twist has been incorporated into the secondary structure determination tool BeStSel (9).

Given the strong correlations between antiparallel sheet distortion and CD spectrum shape (Fig S1), residues assigned as "E" by DSSP are partitioned into three separate classes - P for parallel sheets, AP1 for sheets with minimal distortion and AP2 for "distorted" sheets. Parallel or antiparallel status is determined from information in the DSSP output. Classification of AP1 and AP2 is more complex - in a beta sheet, a residue might interact through hydrogen bonding with multiple residues located in strands N- and C- terminal to its own strand. It might be in a distorted arrangement with one, both or none of its interacting partners. Therefore, we calculate the number of non-distorted vs distorted interactions and apply that ratio to the total count of anti-parallel residues to obtain a final count for the AP1 and AP2 classes.

To determine if a residue, *i*, is part of a distorted strand interaction, the sheet hydrogen bonding network of all anti-parallel residues in the protein is extracted from the DSSP output. The distance and direction between the Cα atom of residue *i* and the Cα atom of the residue N-terminal to *i* in the strand is calculated and stored as a vector as nodes in the network. Edges for node *i* are made to nodes corresponding to residues that form a hydrogen bond with *i.*

All edges in the network are then iterated through and the angle θ between the two vectors *x* and *y* associated with connected nodes, corresponding to the local geometry of their respective strands, is calculated using the dot product:

$$\theta = \cos^{-1}\left(\frac{x \cdot y}{||x||\,||y||}\right)$$

With a θ in the range 35° - 110° the residue is considered to be engaging in a distorted interaction with its partner, and we add one to the count of distorted interactions. Once all interactions have been assessed, we convert the count into a fraction by dividing by the total number of edges in the network. We determine the final total for AP1 and AP2 as below:

$$AP1 = AP - AP2$$
$$AP2 = AP \cdot f_{dist}$$

Where AP is the total number of anti-parallel residues and f_{dist} is the fraction of distorted interactions observed.

Prediction of spectra

Two separate models are used to predict the final spectrum, detailed below.

Least squares model

The secondary structure assignments from DSSP are modified to produce a seven state description of secondary structure, by combining the I ("π helix"), S ("Bend") and B ("isolated β Bridge") assignments into the O class. This results in H1, H2 ("3_{10} helix"), A1, A2, P, T and O. Linear least-squares regression using the secondary structure assignments and circular dichroism spectra of the reference proteins were used to build a model for each wavelength from 180 nm to 260 nm. The *m* CD spectra of all proteins in the reference data set, covering *n* observations at wavelengths in the previously stated range, were stored as an *m* by *n* matrix. This matrix was then transposed to obtain an *n* by *m* matrix with each row containing the CD signal for each protein at a specific wavelength. The secondary structure predictions for each protein were stored as an *m* by 7 matrix. The model weightings (*x*) for each wavelength *n* were obtained by solving the equation $a\,x = b$ by computing a vector *x* that minimizes the Euclidean norm, L^2:

$$L^2 = ||b - ax||^2$$

where *a* is the *m* length vector of CD data points at wavelength *n* and *b* is the matrix containing the secondary structure predictions. From these models, seven basis spectra were calculated by multiplying the weightings *x* at each wavelength *n* by a 7-length vector corresponding to 100% of each secondary structure type in turn.

The prediction is made by multiplying the basis spectra by the corresponding fraction of secondary structure states determined for the query protein by DSSP and summing the resulting spectra.

Linear Combination method

For this method, an eight-state secondary structure assignment is used, adding the DSSP B and S class to the O class, to generate H1, H2, AP1, AP2, P, I, T and O. The 12 proteins with the closest

secondary structure content to the query are selected from the reference data set to act as basis spectra for the prediction. The equation below is then solved using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimisation algorithm implemented in the SciPy python package (10) to find the vector *x* of length *n* corresponding to the weightings applied to the *m* by 8 matrix of the basis protein's secondary structure, b_{pred}, to best fit the secondary structure vector of the query, b_{query}.

$$b_{query} \approx x \cdot b_{pred}$$

The product of the calculated weightings and CD spectra of the *m* basis set proteins results in the predicted CD spectrum of the query.

Calculation of final predicted spectrum.

The spectra resulting from the least squares model and the linear combination method are combined through averaging. The averaged spectrum that results led to better overall accuracy on our reference set (0.996 (0.41)) than either method in isolation (least squares - 0.977 (0.39) Δε, linear combination - 0.996 (0.41) Δε).

Method validation

Leave-one-out cross validation on the training reference data set was used to provide insight on how the method would generalise to an unknown data set. A separate validation was performed on an independent test set consisting of 8 proteins (see Test Set). In all cases the root mean squared deviation (RMSD) of the experimental spectrum versus the predicted spectrum was used to assess accuracy.

The RMSD was calculated using the following equation:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(e_i - p_i)^2}{N}}$$

Where *N* is the number of data points, e_i is the experimental CD signal at each wavelength *i* and p_i is the predicted CD signal at each wavelength *i*.

K-means analysis tool

If the number of predictions made is >=50, the data are partitioned using the k-means clustering algorithm implemented in the SciPy python packaged (10). k-Means clustering partitions observations into k clusters (where k is defined beforehand) so each observation belongs to the cluster with the nearest mean, by minimising the within cluster variances. For k=1 to k=6 inclusive, calculations are performed such that all predicted spectra are clustered using the Euclidean distance between spectra. As the user chooses the value of k, from a dropdown menu, the clustering page automatically updates

the information shown with the pre-calculated information. The mean spectrum, the representative structure (defined as the structure with predicted spectrum with lowest RMSD to the calculated mean) and the average secondary structure values of each cluster at each value of k are obtained. These are presented to the user in the "Clustering" tab of the results page.

An elbow plot showing k versus the distortion is provided to guide the user in the choice of an appropriate value of k. The distortion is calculated as the mean (non-squared) Euclidean distance between the observations passed and the centroids generated for a given value of k. As a general rule of thumb, the optimal value k will be at the "elbow" i.e. the point after which the distortions begin to decrease in a linear fashion.

"Comparison with experiment" tool

The webserver provides a facility to compare a prediction or set of predictions against an uploaded experimental spectrum. Based on user-defined parameters, the tool will produce a subset of predictions closest to the experimental spectra and provide summary statistics on said set. The uploaded spectrum can be in units of Delta Epsilon or Mean Residue Ellipticity (MRE) - selection of the appropriate option on the results page before upload will automatically convert MRE values to delta epsilon using the following equation:

$$\Delta\varepsilon = \mathrm{MRE}/3298$$

The RMSD between all predictions and the experimental spectrum is calculated and the values are then sorted from smallest to largest. There are two options available to the user, choosing subsets of predictions using either a maximum RMSD from experimental spectrum threshold, or by defining a maximum number of predictions, N, to retrieve. Using the former, all predictions with RMSD less than or equal to this value are collected and presented to the user as a set. Using the latter, the N predictions closest to the experimental (as determined by RMSD) forms the set presented to the user.

A range of statistics about the subset are presented to the user, including average secondary structure percentages; RMSD of closest prediction to experimental; RMSD of subset average prediction to the experimental; number of members in the subset; etc. A downloadable .csv file containing the predicted spectra and names of all structures in the set is available to facilitate further analysis by the user.

Cast Study - Unfolding Simulations of Hen Egg White lysozyme

The structure of Hen Egg White lysozyme (HEWL) was obtained from the PDB (PDBID: 2VB1). The CHARMM-GUI webserver (11, 12) was used to generate input for simulations at 270K, 300K, 350K and 450K. All simulations were initiated with 2VB1 as their starting structure. The starting structure was solvated in TIP3 water with potassium and sodium ions added to neutralise the charge of the system. Minimisation, equilibration and production runs were carried out using GROMACS 2019 (13). Minimisation was accomplished using the steepest descent algorithm for 5000 steps. All systems were equilibrated at their specific temperatures for 2.5 ns with a 1 fs timestep using the Nose-Hoover thermostat. Production runs were carried out for 500 ns with a 2 fs timestep using the Nose-Hoover

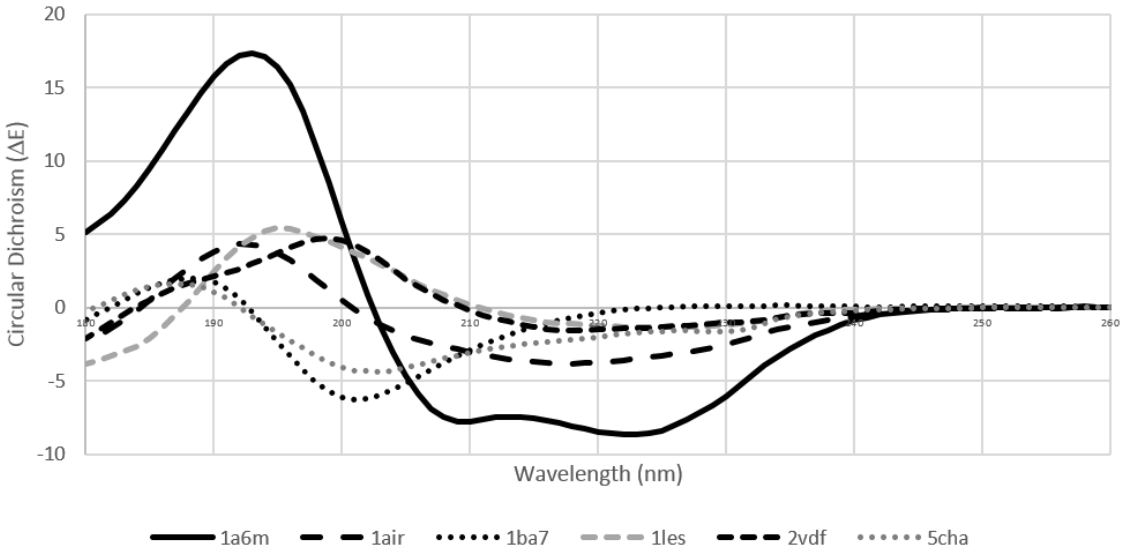
thermostat and Parrinello-Rahman pressure coupling. At each temperature, two simulations were performed yielding 4 μs of total simulation time across all 8 simulations.

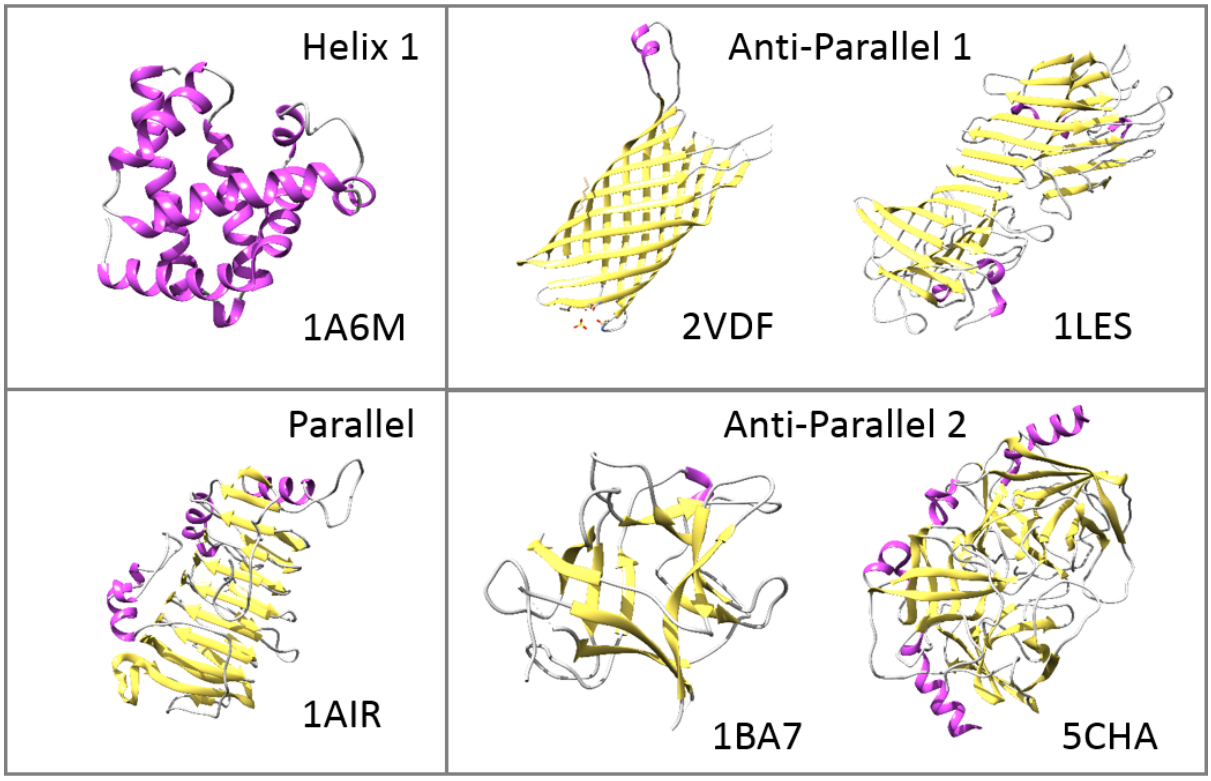
Analysis of MD simulations and comparison with experimental data

Frames were extracted every 1.5 ns from the 8 production runs and saved as PDB formatted files, for a total of 2672 structures that comprised the pool for PDBMD2CD. DSSP was used to obtain the per-residue secondary structure assignments and total secondary structure class count for each structure. The gmx gyrate tool included with GROMACS was used to obtain radius of gyration values for all structures. Predicted spectra for each structure were obtained using the PDBMD2CD webserver.

In vitro experimental CD spectra of HEWL were obtained between 20 ^{O}C and 77 ^{O}C from their PCDDB entries (main text) as was the refold spectrum to 20 ^{O}C from 72 ^{O}C (R72) while data for the 77 ^{O}C (R77) was provided by Dr A.J. Miles (personal communication). Radius of gyration values from Small Angle X-Ray Scattering (SAXS) experiments between 20 ^{O}C and 80 ^{O}C were obtained from data in Meersman et al. (14).

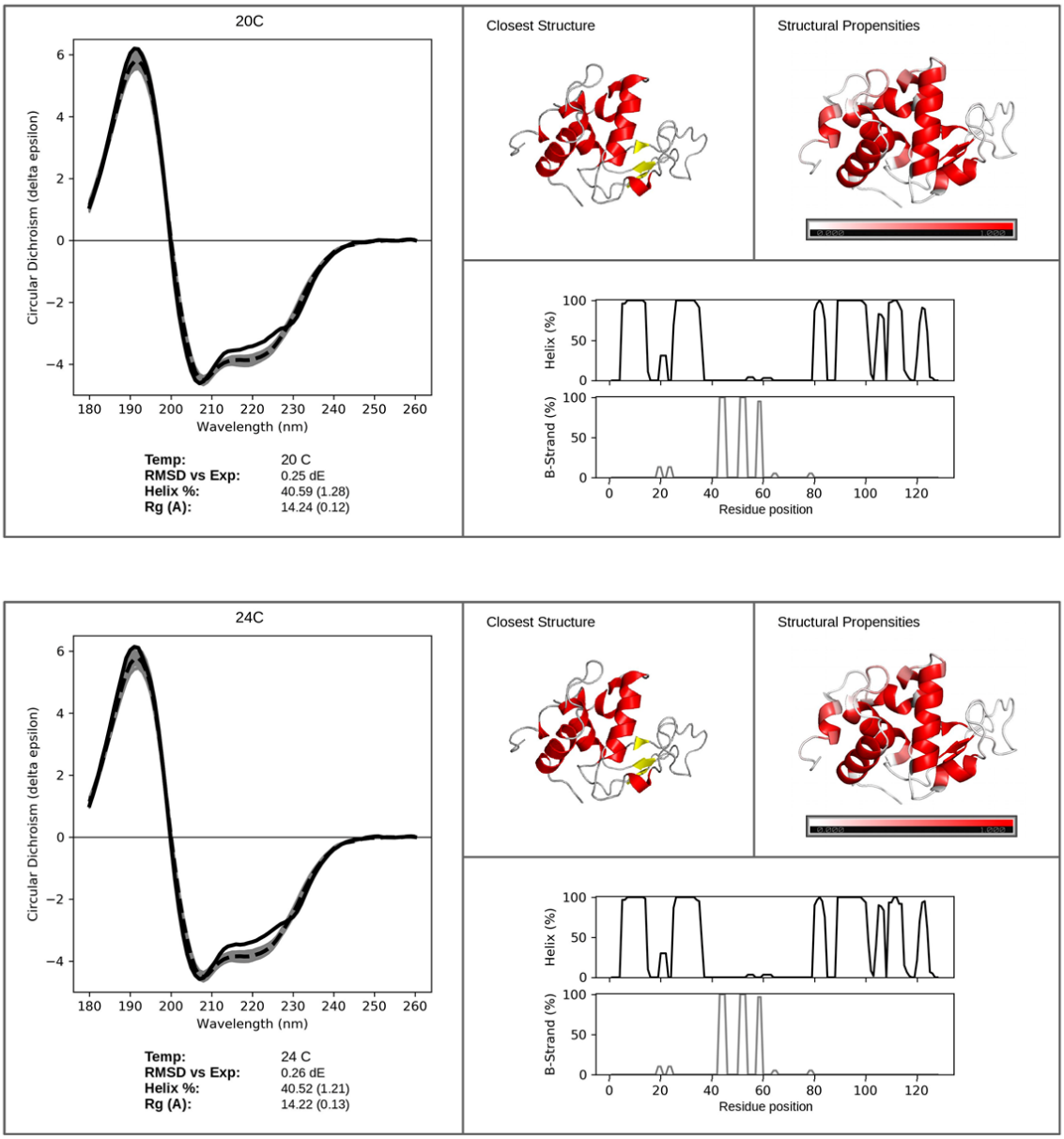
For each experimental CD spectrum, the 60 closest representative structures were obtained using the "Comparison with experiment" tool on the PDBMD2CD website based on the RMSD between the experimental spectra and the predictions. The list of protein structures and their associated predicted spectra were downloaded from the site. Average DSSP secondary structure assignments, per-residue helical propensities and average radius of gyration were calculated for each representative set, and compared against the corresponding data reported in Meersman et al. (14) for the given experimental temperature.

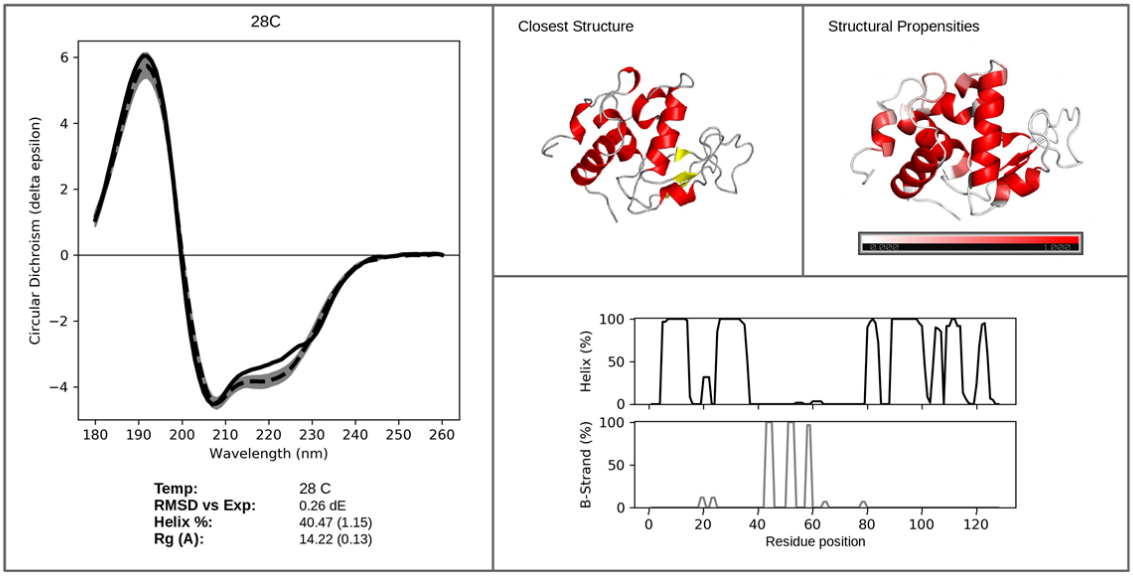


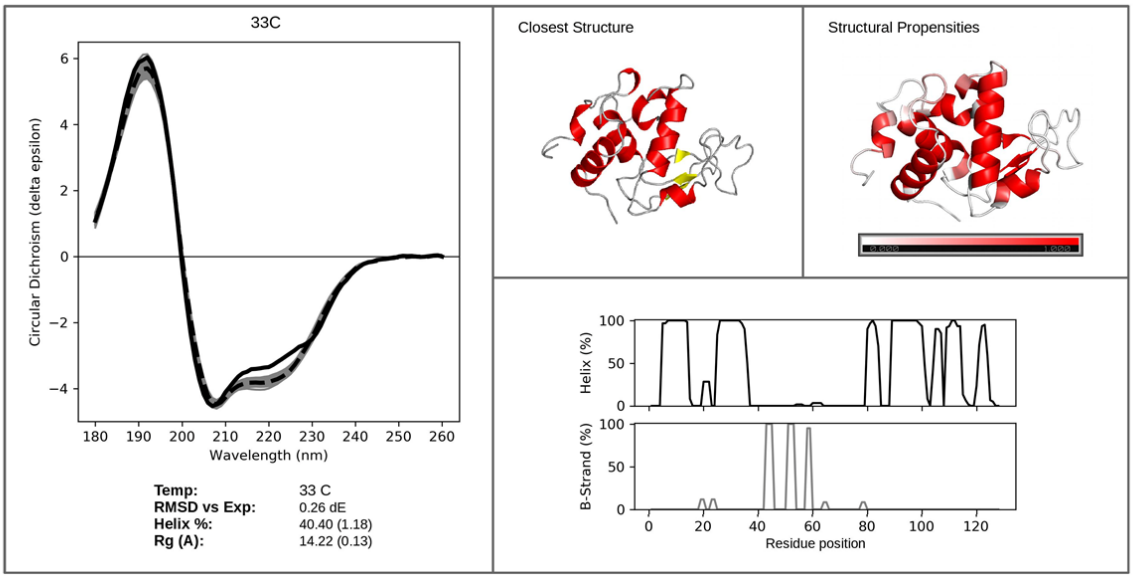


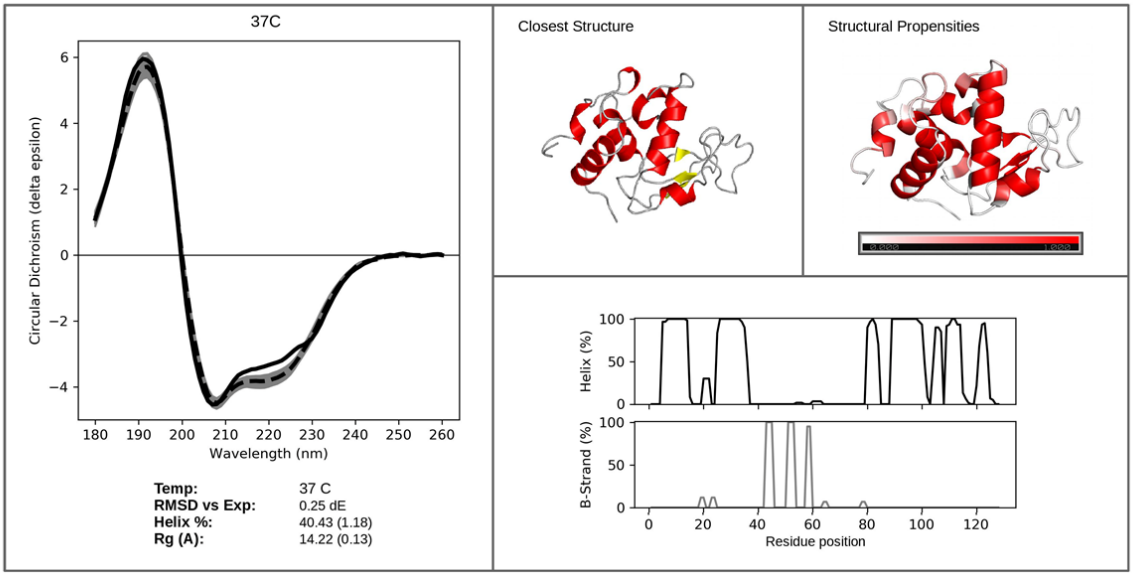
PDB ID	1a6m	1air	1ba7	1les	2vdf	5cha
PCDDBID	CD0000048000	CD0000054000	CD0000065000	CD0000043000	CD0000119000	CD0000005000
Helix 1	0.74	0.09	0.00	0.02	0.02	0.09
Helix 2	0.04	0.04	0.02	0.03	0.00	0.03
Anti-Parallel Strand 1	0.00	0.00	0.15	0.39	0.59	0.16
Anti-Parallel Strand 2	0.00	0.01	0.21	0.08	0.13	0.15
Parallel Strand	0.00	0.31	0.00	0.01	0.02	0.02
Turn	0.09	0.08	0.11	0.11	0.03	0.12
Other	0.13	0.47	0.51	0.37	0.21	0.44

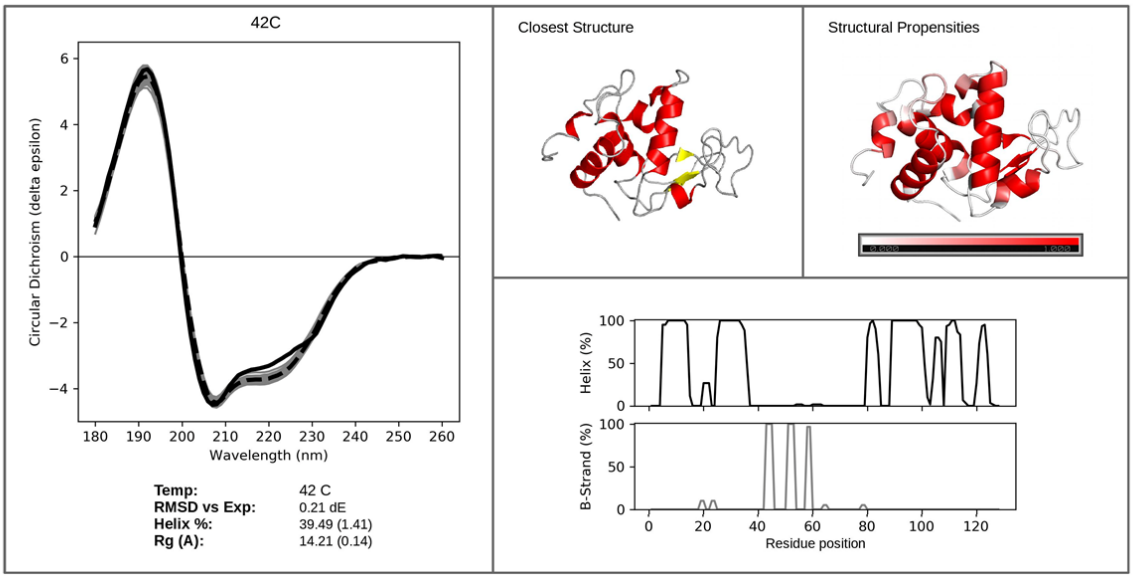
Figure S1: Secondary structure content including beta distortion defines CD spectral characteristics. Presented are the CD spectra, 3D structures and 7 state PDBMD2CD secondary structure classification for six proteins. These act as exemplars for Helix 1 (1A6M – sperm-whale myoglobin), Anti-parallel 1 (minimal distorted) (2VDF - OpcA adhesion protein; 1LES – lentil lectin), Anti-parallel 2 (distorted sheet) (1BA7 - soybean trypsin inhibitor; 5CHA – alpha-chymotrypsin) and parallel (1AIR – pectate lyase C). There is a clear difference between a high content AP1 spectrum which have maxima between ~190 nm and ~200 nm and a high content AP2 spectrum, with a maxima ~186 nm and minima ~202nm. This difference is correlated with the fraction of anti-parallel beta sheet residues involved in "distorted" interactions with their inter-strand hydrogen bonding partners, that has also been noted by others (9).

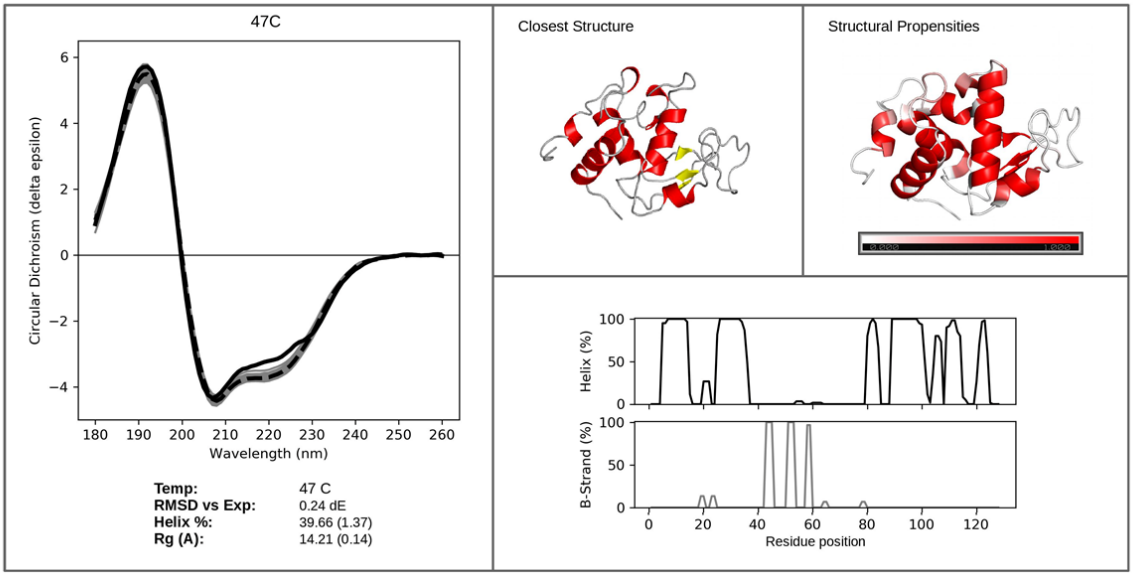


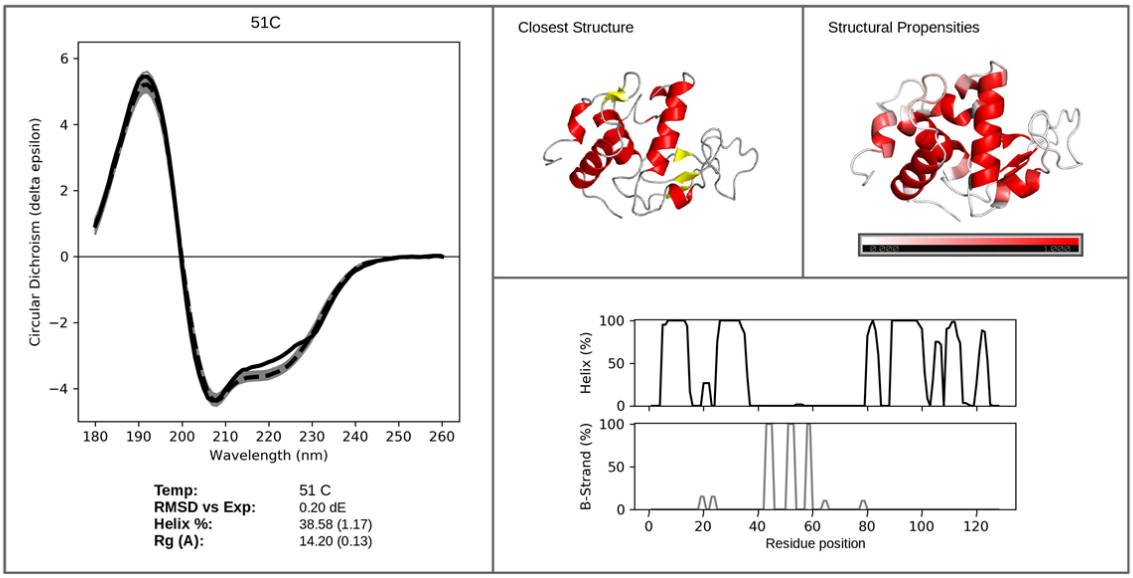

28C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Temp: 28 C
RMSD vs Exp: 0.26 dE
Helix %: 40.47 (1.15)
Rg (A): 14.22 (0.13)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position

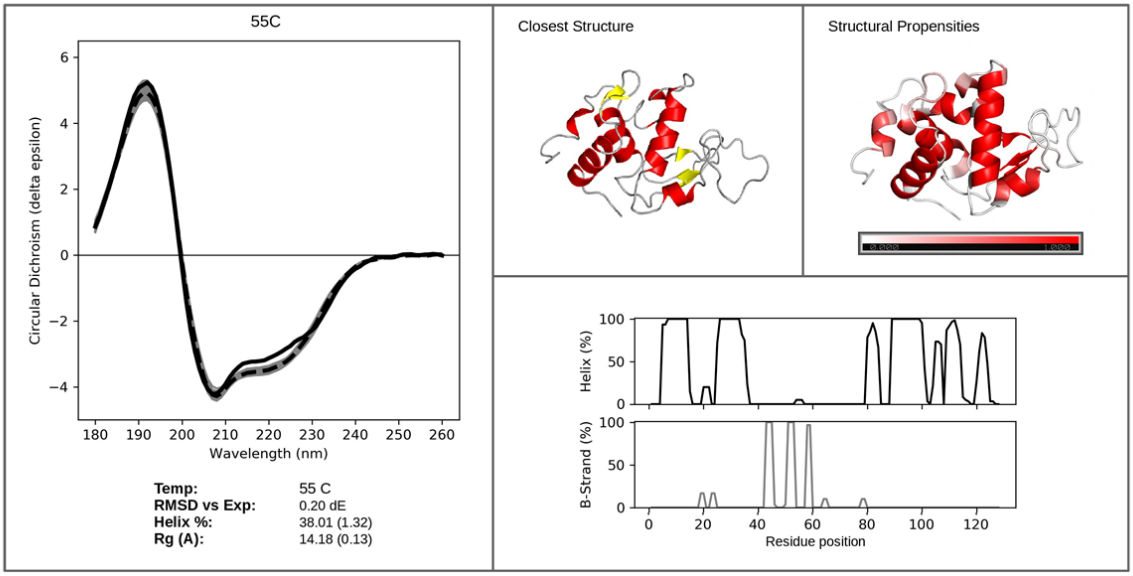

33C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Temp: 33 C
RMSD vs Exp: 0.26 dE
Helix %: 40.40 (1.18)
Rg (A): 14.22 (0.13)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position

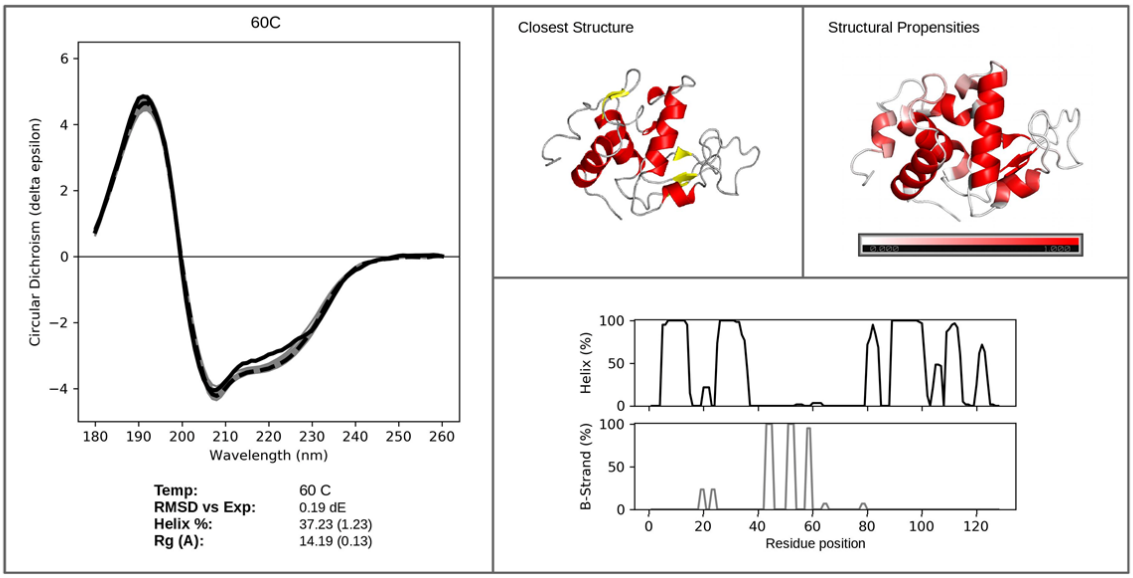

37C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position
Temp: 37 C
RMSD vs Exp: 0.25 dE
Helix %: 40.43 (1.18)
Rg (A): 14.22 (0.13)

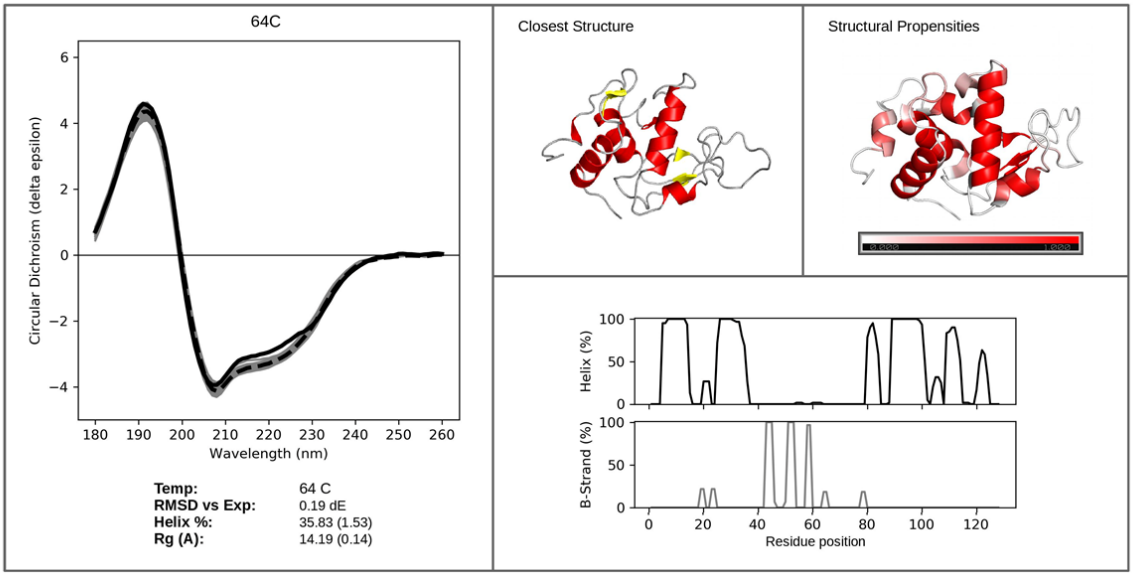

42C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position
Temp: 42 C
RMSD vs Exp: 0.21 dE
Helix %: 39.49 (1.41)
Rg (A): 14.21 (0.14)

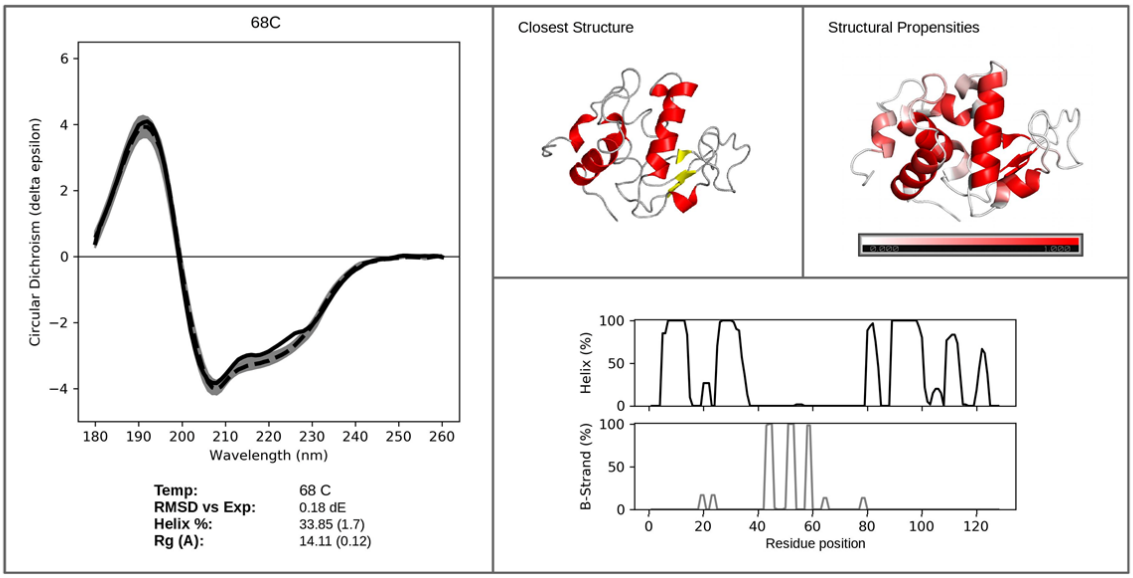

47C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Temp: 47 C
RMSD vs Exp: 0.24 dE
Helix %: 39.66 (1.37)
Rg (A): 14.21 (0.14)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position

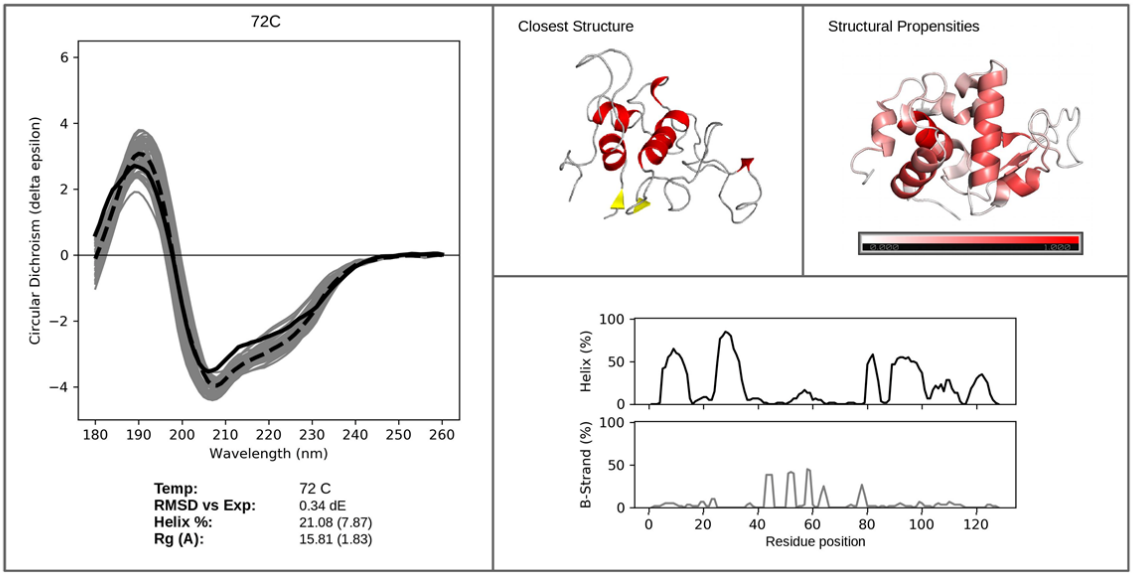

51C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Temp: 51 C
RMSD vs Exp: 0.20 dE
Helix %: 38.58 (1.17)
Rg (A): 14.20 (0.13)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position

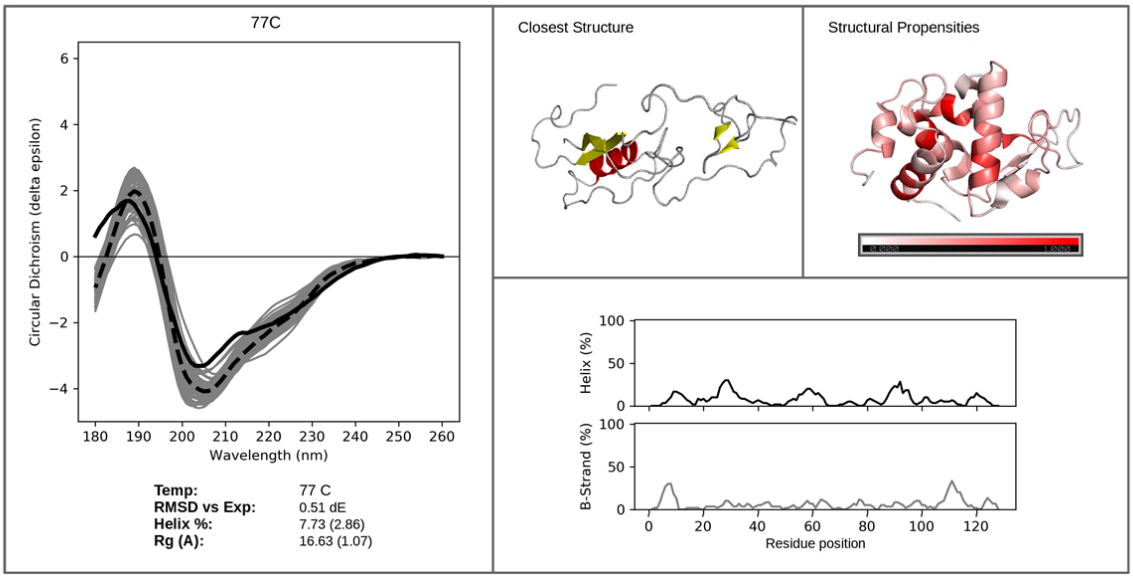

55C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position
Temp: 55 C
RMSD vs Exp: 0.20 dE
Helix %: 38.01 (1.32)
Rg (A): 14.18 (0.13)


60C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position
Temp: 60 C
RMSD vs Exp: 0.19 dE
Helix %: 37.23 (1.23)
Rg (A): 14.19 (0.13)


64C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Temp: 64 C
RMSD vs Exp: 0.19 dE
Helix %: 35.83 (1.53)
Rg (A): 14.19 (0.14)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position


68C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Temp: 68 C
RMSD vs Exp: 0.18 dE
Helix %: 33.85 (1.7)
Rg (A): 14.11 (0.12)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position


72C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Temp: 72 C
RMSD vs Exp: 0.34 dE
Helix %: 21.08 (7.87)
Rg (A): 15.81 (1.83)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position


77C
Circular Dichroism (delta epsilon)
Wavelength (nm)
Temp: 77 C
RMSD vs Exp: 0.51 dE
Helix %: 7.73 (2.86)
Rg (A): 16.63 (1.07)
Closest Structure
Structural Propensities
Helix (%)
B-Strand (%)
Residue position

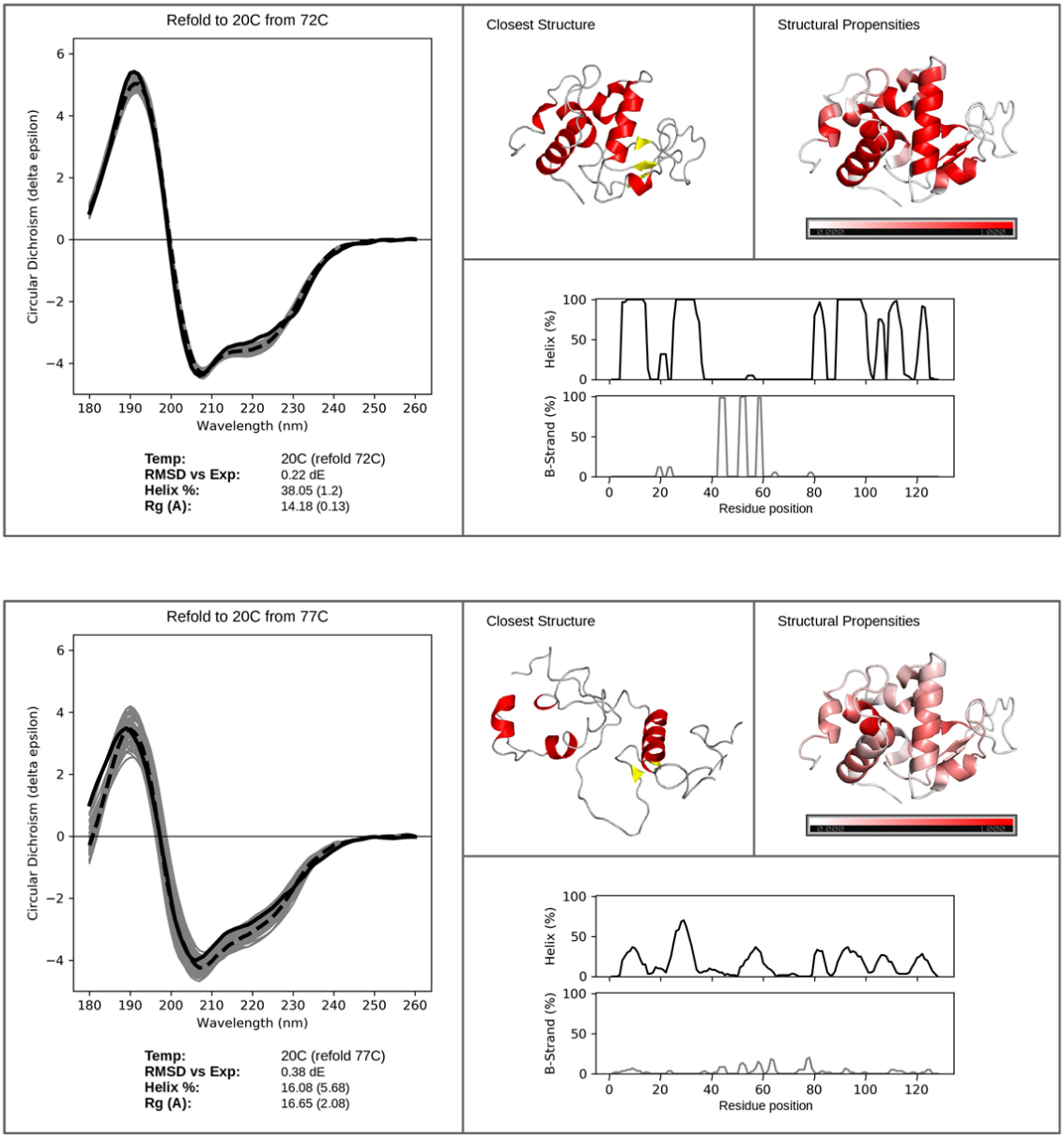


Figure S2: Representative sets produced from fitting of PDBMD2CD predictions of MD-derived structures of lysozyme to experimental CD spectra from 20 ^{O}C to 77 ^{O}C, with refolds to 20 ^{O}C from 72 ^{O}C and 77 ^{O}C. For each temperature, the structures that produced the 60 closest predictions to the experimental spectra (assessed by RMSD) formed a representative set of conformations. A plot of the experimental spectra (black, solid) compared to the 60 predictions (grey) and the average prediction (black, dashed) is shown on the left of each panel. Below the CD plot, the temperature, the RMSD of the average prediction vs the experimental spectra, the average helix percentage of the set and the average radius of gyration (Rg) of the set is detailed. The structure that produced the closest prediction in each set is shown as a "representative structure" for the subset. On the top-right is the native structure of hen egg-white lysozyme from PDB entry 2VB1 coloured according to its per residue, combined helical and beta sheet propensity – this is represented by a gradient from white to red, where white indicates

0% structural propensity for that residue in the set, red 100%. Finally, two plots showing per-residue helix (top) and strand (bottom) propensity can be seen in the bottom right.

PDB ID	PCDDB ID		
1ed9	CD0000002000	1fa2	CD0000009000
1nls	CD0000020000	1m8u	CD0000025000
193l	CD0000045000	3est	CD0000031000
1elp	CD0000024000	1ba7	CD0000065000
2gif	CD0000100000	1dot	CD0000051000
1a49	CD0000061000	1hzx	CD0000123000
2cga	CD0000006000	2nop	CD0000099000
1blf	CD0000042000	1qfe	CD0000028000
1dgf	CD0000017000	1rh5	CD0000124000
3pmg	CD0000057000	1cbj	CD0000068000
3pgk	CD0000058000	7tim	CD0000070000
2dhq	CD0000029000	1ado	CD0000001000
3rn3	CD0000063000	1ppn	CD0000052000
1b8e	CD0000011000	2wjn	CD0000122000
1ubi	CD0000071000	1les	CD0000043000
1nek	CD0000126000	1lin	CD0000013000
2bb2	CD0000022000	1ova	CD0000050000
1xl4	CD0000111000	1ymb	CD0000047000
1t5s	CD0000125000	3dni	CD0000030000
1fcp	CD0000108000	2oar	CD0000115000
5cha	CD0000005000	1une	CD0000059000
1bgl	CD0000010000	1igt	CD0000039000
1l7v	CD0000103000	1hda	CD0000037000
1air	CD0000054000	1fep	CD0000107000
3jqo	CD0000120000	2a65	CD0000113000
1cf3	CD0000033000	2psg	CD0000055000
1a6m	CD0000048000	1hrc	CD0000021000
2j58	CD0000128000	1ax8	CD0000044000
1rhs	CD0000062000	1j95	CD0000110000
1be3	CD0000105000	1nkz	CD0000114000
1hnn	CD0000060000	1a0s	CD0000127000
1bn6	CD0000036000	2vdf	CD0000119000
1ofs	CD0000053000	1hc9	CD0000004000
1ha4	CD0000026000	1kcw	CD0000018000
1hk0	CD0000027000	1k6j	CD0000049000
1thw	CD0000069000	2nr9	CD0000109000
1gpb	CD0000035000	1nqh	CD0000102000
1pcr	CD0000121000	4gcr	CD0000023000

2cts	CD0000019000	1ha7	CD0000012000
1n5u	CD0000038000		
1kpk	CD0000104000		
2cfq	CD0000112000		
2dyr	CD0000106000		
1qhj	CD0000101000		

Table S1: The 83 proteins in the reference set. Shown are the PDB IDs and Protein Circular Dichroism Data Bank (PCDDB) IDs of each reference set protein used to train PDBMD2CD. The spectra for each protein can be found by searching for its ID on the PCDDB website.

DSSP	**Least squares**	**Linear combination**
H	H1	H1
G	H2	H2
I	O	I
E	AP1, AP2, P	AP1, AP2, P
T	T	T
S	O	O
B	O	B
No Class/C	O	O

Table S2: Mapping of DSSP classes to the classifications used in the two predictive models used in PDBMD2CD.

PDB ID	PCDDB ID	Protein Name
1q5u	CD0003897000	human dUTPase
2y3z	CD0003898000	3-isopropylmalate dehydrogenase
1qlp	CD0003890000	Alpha-1-antitrypsin
2ccm	CD0004676000	Calexcitin
1sr5	CD0003889000	Antithrombin-III
4kyp	CD0004244000	Bj-xtrIT
1ecz	CD0003896000	Ecotin
2yxf	CD0003894000	Beta-2-microglobulin

Table S3: The eight proteins used as the Test set for this package as chosen from the PCDDB (3)

SUPPLEMENTARY REFERENCES

1. Abdul-Gader,A., Miles,A.J. and Wallace,B.A. (2011) A reference dataset for the analyses of membrane protein secondary structures and transmembrane residues using circular dichroism spectroscopy. *Bioinformatics,* 27, 1630-1636. https://doi.org/10.1093/bioinformatics/btr234

2. Whitmore, L. and Wallace, B.A. (2008) Protein Secondary Structure Analyses from Circular Dichroism Spectroscopy: Methods and Reference Databases. *Biopolymers*, 89, 392-400. (PDF)

3. Whitmore,L., Miles,A.J., Mavridis,L., Janes,R.W. and Wallace,B.A. (2017) PCDDB: new developments at the Protein Circular Dichroism Data Bank. *Nucleic Acids Res*., 45(D1), D303-D307. http://nar.oxfordjournals.org/content/45/D1/D303

4. Woollett,B., Whitmore,L., Janes,R.W. and Wallace,B.A. (2013) ValiDichro: a website for validating and quality control of protein circular dichroism spectra. *Nucleic Acids Res*., 41(W1), W417–W421. https://doi.org/10.1093/nar/gkt287

5. Lees,J.G., Miles,A.J., Wien,F. and Wallace,B.A. (2006) A reference database for circular dichroism spectroscopy covering fold and secondary structure space. *Bioinformatics,* 22, 1955-1962. http://www.ncbi.nlm.nih.gov/pubmed/16787970

6. Kabsch.W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22, 2577-637. doi:10.1002/bip.360221211. PMID 6667333.

7. Manavalan,P. and Johnson,W. (1983) Sensitivity of circular dichroism to protein tertiary structure class. *Nature*, 305**,** 831–832. https://doi.org/10.1038/305831a0

8. Wu,J., Yang,J.T. and Wu,C.S.C. (1992) β-II conformation of all-β proteins can be distinguished from unordered form by circular dichroism. *Analytical Biochem*., 200, 359-364. https://doi.org/10.1016/0003-2697(92)90479-Q.

9. Micsonai,A., Wien,F., Bulyáki,E., Kun,J., Moussong,E., Lee,Y.H., Goto,Y., Réfrégiers,M. and Kardos,J. (2015) Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy, *Proc. Nat. Acad. Sci.*, 112, E3095-E3103. doi/10.1073/pnas.1500851112

10. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J., van der Walt,S.J., Brett,M., Wilson,J., Millman,K.J., Mayorov,N., Nelson,A.R.J., Jones,E., Kern,R., Larson,E., Carey,C.J., Polat,İ., Feng,Y., Moore,E.W., VanderPlas,J., Laxalde,D., Perktold,J., Cimrman,R., Henriksen,I., Quintero,E.A., Harris,C.R., Archibald,A.M.,

Ribeiro,A.H., Pedregosa,F., van Mulbregt,P. and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261-272. doi.org/10.1038/s41592-019-0686-2

11. Jo,S., Kim,T., Iyer,V.G. and Im,W. (2008) CHARMM-GUI: A Web-based Graphical User Interface for CHARMM. *J. Comput. Chem*., 29, 1859-1865. doi.org/10.1002/jcc.20945

12. Lee,J., Cheng,X., Swails,J.M., Yeom,M.S., Eastman,P.K. Lemkul,J.A., Wei,S., Buckner,J., Jeong,J.C., Qi,Y., Jo,S., Pande,V.S., Case,D.A., Brooks III,C.L., MacKerell Jr,A.D., Klauda,J.B. and Im,W. (2016) CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput*., 12, 405-413. doi.org/10.1021/acs.jctc.5b00935

13. Abraham,M.J., Murtola,T., Schulz,R., Páll,S., Smith,J.C., Hess,B. and Lindahl,E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX*, 1–2, 19-25. https://doi.org/10.1016/j.softx.2015.06.001.

14. Meersman,F., Atilgan,C., Miles,A.J., Bader,R., Shang,W.F., Matagne,A., Wallace,B.A. and Koch,M.H.J. (2010) Consistent Picture of the Reversible Thermal Unfolding of Hen Egg-White Lysozyme from Experiment and Molecular Dynamics. *Biophysical J.,* 99, 2255-2263. https://doi.org/10.1016/j.bpj.2010.07.060